When Heavy-Tailed and Light-Tailed Flows Compete: The Response Time Tail Under Generalized Max-Weight Scheduling

Jayakrishnan Nair[1], Krishna Jagannathan[2], and Adam Wierman[1]

[1]Computing and Mathematical Sciences, California Institute of Technology, USA
[2]Department of Electrical Engineering, IIT Madras, India

***Abstract*—This paper focuses on the design and analysis of scheduling policies for multi-class queues, such as those found in wireless networks and high-speed switches. In this context, we study the response time tail under generalized max-weight policies in settings where the traffic flows are highly asymmetric. Specifically, we study an extreme setting with two traffic flows, one heavy-tailed, and one light-tailed. In this setting, we prove that classical max-weight scheduling, which is known to be throughput optimal, results in the light-tailed flow having heavy-tailed response times. However, we show that via a careful design of inter-queue scheduling policy (from the class of generalized max-weight policies) and intra-queue scheduling policies, it is possible to maintain throughput optimality, and guarantee light-tailed delays for the light-tailed flow, without affecting the response time tail for the heavy-tailed flow.**

I. Introduction

The task of scheduling conflicting links is central to a variety of networking settings, such as wireless networks, optical networks and high-speed switches. As a result, there is a large literature studying scheduling policies in these contexts, most of which is based on the maximum-weight (max-weight) scheduling framework proposed by Tassiulas and Ephremides in [1], [2]. At this point, there is a substantial body of literature devoted to the analysis and application of the max-weight policy and its variants; for example, see [3]–[9].

Traditionally, the focus of research on max-weight scheduling has been on understanding its 'stability region', i.e., the set of input rates that can be supported. Notably, max-weight has been shown to be 'throughput optimal' in very general settings, i.e., it has the largest possible stability region among all scheduling policies [1], [2], [9]. In other words, if there exists any scheduling policy that can keep the queueing network stable under a given model of traffic arrival statistics, the max-weight policy can stabilize the system.

Although throughput is an important first-order performance metric, a more discerning metric is the *response time*, a.k.a., sojourn time or delay. Indeed, from the standpoint of the applications sending/receiving information, ensuring small, predictable response times is crucial. Although the stability region and throughput optimality properties of the max-weight framework are well studied, the literature on the delay performance is relatively small. Average delay bounds are derived in [9] using Lyapunov drift techniques; however, these are quite loose in general. Tighter delay bounds do exist, e.g., those in [10] for max-weight scheduling in spatially homogeneous wireless ad-hoc networks.

In general, results about the response time of max-weight policies, such as those above, tend to indicate that max-weight policies perform well in symmetric traffic settings. This is primarily due to the tendency of these policies to 'balance out' the queues in the system, by preferentially serving longer queues. For example, [1] contains a strong sample path optimality result for queue backlogs under stochastically symmetric traffic to parallel queues; this is generalized in [11].

On the other hand, the traffic flows encountered in practice tend to be highly asymmetric, with a wide range of variability or burstiness. Indeed, in the context of communication networks, certain bursty traffic flows may be well modeled using heavy-tailed arrival processes, and the more benign ones better modeled using light-tailed processes. For example, an internet user might generate occasional file download requests with highly variable file sizes, that can be modeled as being heavy-tailed. However, routine webpage loading and email traffic are likely to be far less variable, and thus are better modeled as being light-tailed. In order to capture the interaction between heterogeneous traffic sources in a queueing network, multi-class queueing models with a mix of heavy-tailed and light-tailed traffic sources have been studied [12]–[14]. An important paper in this category is [12], where the interaction between light and heavy-tailed traffic flows under generalized processor sharing (GPS) is studied. Another example is [13], where the authors obtain the asymptotic workload behavior under a general coupled-queues framework, which includes GPS as a special case.

So, on the one hand, max-weight policies are throughput optimal and provide good response times when the traffic is largely symmetric. On the other hand, the interaction between bursty and benign traffic sources is well studied within multi-class queueing and GPS frameworks, but these policies are not throughput optimal.

Contributions of this paper

The goal of this paper is to fill this gap by studying response times under max-weight policies when traffic is highly asymmetric. The first steps towards filling this gap have been provided by the recent work of Markakis et al. in [15] and Jagannathan et al. [16], [17], which analyze a scenario where heavy-tailed and light-tailed flows interact through a generalized max-weight policy. Our present paper builds on these papers; in particular, our model is borrowed from [16]. However, the focus of the above papers is on *queue length* asymptotics under different throughput optimal policies, while in this paper, we analyze the distribution of *response times* experienced by the heavy and light-tailed flows.

More specifically, in this paper, we consider a stylized setting where the traffic asymmetry is extreme. We consider a system consisting of two traffic classes contending for service from a single server, where one class is heavy-tailed, and the other is light-tailed (see Fig. 1). Both classes experience a time varying connectivity with the server, and the server can serve a single packet from a connected queue in each slot. Note that this model captures a wireless uplink/downlink scenario with two nodes communicating with an access point or base station via fading channels. For this queueing system, we study the tail of the (stationary) response time distribution each traffic class experiences under generalized max-weight policies.

In this context, there are two scheduling decisions: the *inter-queue* scheduling and the *intra-queue* scheduling. The inter-queue scheduling policy determines which queue to serve in each slot, whereas the intra-queue scheduling policies specify which waiting packet to serve from the queue selected for service by the inter-queue scheduling policy.

The *first contribution* of this paper is to prove that the classical max-weight policy, which serves the longest connected queue in each slot, causes the light-tailed flow to experience heavy-tailed response times. This means that the classical max-weight policy, while being throughput optimal, severely throttles (starves) the light-tailed flow. Thus, while max-weight performs well in symmetric settings, it can have poor performance in asymmetric settings. Intuitively, this is because the max-weight policy starves the light-tailed flow of service for a long period of time when the heavy-tailed flow generates its (frequent) large bursts.

The *second contribution* of this paper is to show that it is possible to design a throughput optimal scheduling policy that avoids the problems experienced by the classical max-weight policy. In particular, we present a policy that provably guarantees light-tailed response times for the light-tailed flow. Importantly, our results suggest that the response time tail for the heavy-tailed flow remains unaffected; we prove this formally for the special case in which both queues are always connected to the server.

Our policy design entails a careful choice of the inter-queue scheduling policy, as well as intra-queue scheduling policies. Our inter-queue policy is the so called 'log-max-weight policy', which belongs to the class generalized max-weight policies [6] and awards a significant priority to the light-tailed flow, while maintaining throughput optimality. Our intra-queue policy differs between the heavy-tailed and the light-tailed queues: within the heavy-tailed queue, Preemptive-Last-Come-Fist-Served (PLCFS) is used, while within the light-tailed queue, First-Come-First-Served (FCFS) is used.

Our analysis provides a clear insight into the intricate interplay between the intra-queue and inter-queue scheduling policies. Indeed, our results reveal that even with a good inter-queue scheduling policy, the correct choice of intra-queue scheduling policies is crucial in order to obtain good response time tail behavior. In fact, the difference in response times between two intra-queue policies can be significantly larger under generalized max-weight inter-queue scheduling than in a single server queue.

Finally, it is worth commenting that in attaining the results described above, we also settle an an open question in [17, pp. 171] regarding the asymptotics of log-max-weight scheduling. In particular, we prove that under log-max-weight scheduling, the (stationary) queue length distribution corresponding to the light-tailed queue is light-tailed (Theorem 9), via a novel application of Lyapunov bounds from [6].

II. Model and preliminaries

A. System model

Our goal is to study multi-class queues in a setting where the traffic flows are highly asymmetric. To that end, we consider a simple model where the asymmetry is extreme. In particular, we consider a scenario where two parallel queues contend for service from a single server. One of the queues sees a heavy-tailed arrival process, whereas the other sees a light-tailed arrival process. We refer to the former queue as the heavy queue, and the latter queue as the light queue.

Each queue experiences a stochastically time varying connectivity with the server. Fig. 1 provides an illustration of our setup. Time is slotted, and in each slot, the server can provide a single unit of service to a connected queue. Henceforth, we refer to this unit of service as a packet, and say the server can process a single packet from a connected queue in each slot. Let t denote the time index.

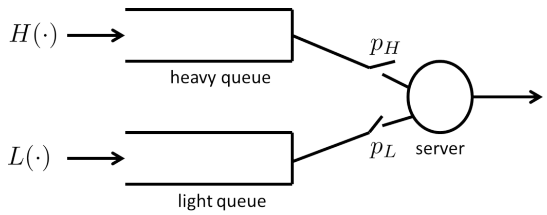

Fig. 1: A network consisting of two parallel queues, with one of them fed with heavy-tailed traffic. The channels connecting the queues to the server are unreliable ON/OFF links.

In each slot, a job, comprising a burst of packets, can arrive stochastically into each queue. Let $H(t)$ and $L(t)$ denote, respectively, the size of the job (in number of packets) arriving into the heavy queue and the light queue in time slot t. We adopt the convention that the size of the incoming job is zero if there is no arrival in a slot.

Our stochastic model for the arrival processes is the following. The sequences $L(\cdot)$ and $H(\cdot)$ are i.i.d. across time slots, and independent of one another. The random variable $L(t)$ is light-tailed, and the random variable $H(t)$ is heavy-tailed. Specifically, we assume that $H(t)$ is regularly varying with index $\theta_H > 1$.[1] Let $\lambda_H := \mathbb{E}[H(t)]$ and $\lambda_L := \mathbb{E}[L(t)]$ denote the mean arrival rates into the heavy queue and the light queue, respectively.

Next, we describe the stochastic model for the connectivity of each queue with the server. The connectivity of the heavy queue and the light queue are described, respectively, by Bernoulli sequences $\{\eta_H(t)\}$ and $\{\eta_L(t)\}$. $\eta_H(t), \eta_L(t) \in \{0, 1\}$, with a value of 1 indicating that the corresponding queue is connected to the server in time slot t. We assume that the sequences $\{\eta_H(t)\}$ and $\{\eta_L(t)\}$ are mutually independent and independent of the arrival processes. Let $p_H := P(\eta_H(t) = 1)$ and $p_L := P(\eta_L(t) = 1)$ denote, respectively, the probabilities that the heavy queue and the light queue are connected to the server in each time slot. We assume that $p_L, p_H > 0$. We refer to the special case

[1] We formally define light-tailed and regularly varying distributions in Section II-D.

of our model in which the two queues are always connected to the server, i.e., $p_L = p_H = 1$, as the *wireline scenario*. For technical reasons, we exclude from consideration the scenario where only one of the queues is always connected to the server, i.e., we exclude the cases $p_L = 1,\ p_H \in (0,1)$ and $p_H = 1,\ p_L \in (0,1)$. Finally, we assume that the server can detect the connectivity state of both queues, as well as the queue size (in number of packets) of a connected queue in each slot. Note that our model captures an uplink/downlink setting with two wireless nodes connected to a base station or access point via independent fading channels.

Let $q_H(t)$ and $q_L(t)$ denote, respectively, the lengths (in number of packets) of the heavy queue and the light queue in the beginning of time slot t. The queue lengths evolve as follows.

$$\begin{aligned} q_H(t+1) &= H(t) + q_H(t) - \mathbf{1}_{\{\text{heavy queue got service in slot t}\}}, \\ q_L(t+1) &= L(t) + q_L(t) - \mathbf{1}_{\{\text{light queue got service in slot t}\}}. \end{aligned}$$

If both queues are connected to the server in a certain slot, the scheduling policy determines which queue will receive service. If only one of the queues is connected to the server in a certain slot, then that queue receives service if it has any waiting packets. We refer to such slots as exclusive slots. We use q_H and q_L to denote, respectively, the stationary queue lengths of the heavy queue and the light queue. We use V_H to denote the steady state response time experienced by a job in the heavy queue, and V_L to denote the steady state response time experienced by a job in the light queue.

B. Stability region

The *stability region* for the queueing system defined above, i.e., the set Λ of (λ_H, λ_L) pairs that are stabilizable, is well understood. It follows from [1] that

$$\begin{aligned} \Lambda = \{(\lambda_H, \lambda_L) \mid 0 \leq & \lambda_H < p_H,\ 0 \leq \lambda_L < p_L, \\ & \lambda_H + \lambda_L < p_H + p_L - p_H p_L\}. \end{aligned}$$

The constraints defining the stability region are intuitive: the average arrival rate into each queue cannot exceed its maximum possible service rate (defined by fraction of time it is connected to the server), and the sum total of the arrival rates cannot exceed the maximum possible aggregate service rate of the two queues (defined by the fraction of time at least one queue is connected to the server). The stability region is visualized in Fig. 2. We seek scheduling policies that are *throughput optimal*, i.e., policies that stabilize the queueing system over the entire stability region.

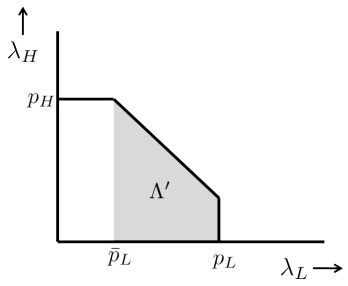

Fig. 2: The stability region Λ is the pentagonal region above. The subset Λ' of interest is shaded.

Let $\bar{p}_L := p_L(1 - p_H)$. Note that $\bar{p}_L$ is the probability that only the light queue is connected to the server in a slot, i.e., the probability that a slot is exclusive to the light queue. If $\lambda_L < \bar{p}_L$, then the arrivals into the light queue can be stably supported by just exclusive slots, implying the light queue essentially does not need to compete for service with the heavy queue. This case is uninteresting when analyzing the light queue, since the response time distribution is guaranteed to be light-tailed, irrespective of the inter-queue or intra-queue scheduling policy. For the same reason, the case $\lambda_H = 0$ is uninteresting. Therefore, when studying the response time distribution in the light queue, we restrict our attention to the subset Λ' of the stability region over which $\lambda_L > \bar{p}_L$, and $\lambda_H > 0$. The set Λ' is depicted in Fig. 2. Note that in the wireline scenario, $\bar{p}_L = 0$, and Λ' is simply the interior of the stability region.

C. Scheduling

We decouple the scheduling design as follows. The *inter-queue scheduling policy* determines which queue to serve in each slot, given the connectivity state and length (in number of packets) of each queue. In the queue selected for service by the inter-queue policy, the *intra-queue scheduling policy* determines which packet to serve in that slot, given the full state of the queue. We consider a variety of possible policies, described below, for each.

Recall that we have two performance goals for scheduler design: (i) throughput optimality, and (ii) good response time tail behavior. Note that the stability of the queueing system depends solely on the inter-queue scheduling policy, since the evolution of the queue lengths is insensitive to the intra-queue scheduling policy. However, the response time distribution is highly dependent on the intra-queue scheduling policy.

1) Inter-queue scheduling: Given that the inter-queue scheduling policy completely determines the stability of the system, it is crucial to use policies that are throughput optimal. This motivates us to consider generalized max-weight policies [6]. In particular, our focus is on two such policies:

Max-weight-α scheduling: The max-weight-α policy [15], [17] is a generalization of the classical max-weight policy, and is characterized by two positive parameters α_L and α_H. In each slot, the max-weight-α policy serves the queue that wins the comparison

$$q_L(t)^{\alpha_L}\eta_L(t) \gtreqless q_H(t)^{\alpha_H}\eta_H(t).$$

Ties may be broken arbitrarily, but we assume for concreteness that ties are broken in favor of the light queue. Note that when $\alpha_L = \alpha_H$, the max-weight-α policy is identical to the classical max-weight policy. The throughput optimality of this policy follows easily from Theorem 1 in [6].

The parameters α_L and α_H determine the relative priorities of the two queues. Since we will be interested in the scenario where the light queue receives a higher priority than the heavy queue, we focus on the case $\alpha_L \geq \alpha_H$. Moreover, it is easy to see that we may set $\alpha_H = 1$ without loss of generality. Accordingly, we focus on the range of parameters satisfying $\alpha_L \geq \alpha_H = 1$. Note that a higher value of α_L implies a higher priority for the light queue.

Log-max-weight scheduling: The log-max-weight policy [17] is defined as follows. In each slot t, it serves the queue that wins the comparison

$$q_L(t)\eta_L(t) \gtreqless \log(1 + q_H(t))\eta_H(t). \tag{1}$$

As before, we assume for concreteness that ties are broken in favor of the light queue. The throughput optimality of this policy once again follows easily from Theorem 1 in [6].

The log-max-weight policy awards an even higher degree of priority to the light queue than the max-weight-α policy. Note that in order to determine which queue to serve in a slot, the max-weight-α policy compares $q_H(t)$ with $q_L(t)^{\alpha_L}$, whereas the log-max-weight policy compares $q_H(t)$ with $e^{q_L(t)} - 1$.

2) Intra-queue scheduling: While intra-queue scheduling does not impact the stability of the system (as long as the policies considered are work-conserving), the intra-queue scheduling policy does have a significant impact on the response time distribution. In this paper, we focus on two candidate policies for intra-queue scheduling: First-Come-First-Served (FCFS) and Preemptive-Last-Come-First-Served (PLCFS).

While other policies could also be considered, the choice of these policies is motivated by a few important factors. First, FCFS is the most commonly assumed intra-queue policy in the literature on max-weight scheduling. Second, there have been suggestions recently that using PLCFS as the intra-queue scheduling policy can improve the delay-performance of max-weight policies [18]. Third, in a single server queue, it is known that the response time tail under FCFS is optimal when job sizes are light-tailed, while the response time tail under PLCFS is optimal (up to a constant) when job sizes are heavy-tailed (see [19]).

D. Heavy-tailed distributions: Definitions and properties

In this section, we give relevant definitions and preliminaries related to heavy-tailed distributions.

For any non-negative random variable X, we use F_X to denote its distribution function (d.f.), i.e., $F_X(x) := P(X \leq x)$, and $\bar{F}_X$ to denote its tail distribution function, i.e., $\bar{F}_X(x) := P(X > x)$. The random variable X (or its d.f. F_X) is said to be *heavy-tailed* if

$$\limsup_{x\to\infty} \frac{\bar{F}_X(x)}{e^{-\phi x}} = \infty \qquad \forall\ \phi > 0.$$

Conversely, X (or its d.f. F_X) is said to be *light-tailed* if it is not heavy-tailed, i.e., if there exists $\phi > 0$ such that

$$\lim_{x\to\infty} \frac{\bar{F}_X(x)}{e^{-\phi x}} = 0.$$

Intuitively, a d.f. is heavy-tailed if its tail is asymptotically heavier than that of any exponential distribution.

An important characterization of heavy-tailed distributions that we make use of in our analysis is the following. See [20, Chap. 5] for a proof of this lemma. For any non-negative random variable X, define $\Psi_X(x) := \frac{-\log \bar{F}_X(x)}{x}$.

Lemma 1. *Suppose X is non-negative random variable. Then X is heavy-tailed if and only if*

$$\liminf_{x\to\infty} \Psi_X(x) = 0.$$

From a modeling standpoint, an important subclass of heavy-tailed distributions is the class of regularly varying distributions, which is a generalization of the class of Pareto distributions [21]. Formally, a random variable X (or its d.f. F_X) is said to be *regularly varying* with index $\theta > 0$ (denoted $X \in \mathcal{RV}(\theta)$) if $P(X > x) = x^{-\theta}L(x)$, where $L(x)$ is a slowly varying function, i.e., $L(x)$ satisfies $\lim_{x\to\infty} \frac{L(xy)}{L(x)} = 1\ \forall\ y > 0$. Recall that our model assumes that $H(t) \in \mathcal{RV}(\theta_H)$.

Our focus in this paper is on understanding the (logarithmic) asymptotic behavior of the response time tail. To study this for a heavy-tailed X, we use its *tail index*, defined as

$$\Gamma(X) := \lim_{x\to\infty} -\frac{\log P(X > x)}{\log(x)},$$

when the limit exists. The tail index is useful for describing the asymptotic tail behavior of distributions that exhibit a roughly 'power-law' tail, such as regularly varying distributions. In particular, if $X \in \mathcal{RV}(\theta)$, then $\Gamma(X) = \theta$ [22, Prop. 2.6]. It is easy to check that if $\Gamma(X) < \infty$, then X is heavy-tailed. Moreover, it can be shown that

(i) if $\Gamma(X) > 0$, then $\mathbb{E}\left[X^\beta\right] < \infty$ for $0 \leq \beta < \Gamma(X)$,
(ii) if $\Gamma(X) < \infty$, then $\mathbb{E}\left[X^\beta\right] = \infty$ for $\beta > \Gamma(X)$.

Finally, note that a smaller value of tail index implies a 'heavier' tail.

To give a lower bound on the tail of a heavy-tailed random variable X, we use

$$\bar{\Gamma}(X) := \limsup_{x\to\infty} -\frac{\log P(X > x)}{\log(x)}.$$

It is easy to check that if $\bar{\Gamma}(X) < \infty$, then X is heavy-tailed. Moreover, if $\bar{\Gamma}(X) < \infty$, then $\mathbb{E}\left[X^\beta\right] = \infty$ for $\beta > \bar{\Gamma}(X)$.

III. Results

The multi-class queueing model in the previous section involves two highly asymmetric traffic classes, one heavy-tailed, and one light-tailed. Our goal now is to understand how max-weight scheduling and its variants perform under such an extreme form of asymmetry. We begin by considering the most well studied subclass of generalized max-weight policies: max-weight-α policies [15], which includes the classical max-weight policy as a special case. We then consider the class of log-max-weight policies.

Recall that both of these classes of inter-queue policies ensure throughput stability regardless of the intra-queue policy used. Therefore, our results focus on the response time tail. Importantly, for this metric, our results highlight that the choice of intra-queue scheduling is crucial.

A. Max-weight-α scheduling

In this section, we present our results on the tail behavior of the (stationary) response time distribution in the heavy queue and light queue under the max-weight-α inter-queue scheduling policy. We being by focusing on the light-queue.

The performance of the light queue: Our first result is the following upper bound on the response time tail index for the light queue under max-weight-α inter-queue scheduling, and any intra-queue scheduling policy in the light queue.

Theorem 2. *Suppose that the arrival rates lie in the subset Λ' of the stability region. Then under the max-weight-α scheduling policy between queues with $\alpha_L \geq \alpha_H = 1$,*

$$\bar{\Gamma}(V_L) = \limsup_{x\to\infty} -\frac{\log P(V_L > x)}{\log(x)} \leq \alpha_L\theta_H - 1$$

for any intra-queue scheduling policy in the light queue.

Theorem 2 states that under max-weight-α scheduling between queues, $\bar{\Gamma}(V_L) < \infty$, which implies that *the light*

queue sees heavy-tailed response times, irrespective of the intra-queue scheduling policy. This means that although max-weight-α scheduling is throughput optimal, it severely throttles the light queue. Note that this includes the classical max-weight policy as a special case. Intuitively, this poor performance is the result of (frequent) large arrivals into the heavy queue starving the light queue of service for a long time.

However, it is important to note that the upper bound on the response time tail index given by Theorem 2 is an increasing function of α_L, approaching ∞ as $\alpha_L \to \infty$. This suggests the possibility of achieving an arbitrarily large response time tail index for the light queue (recall that a larger tail index implies a lighter tail) by setting α_L large enough, i.e., by awarding the light queue sufficiently high priority. Theorems 3 and 4 below imply that this is indeed the case, so long as the intra-queue policy in the light queue is chosen appropriately. Intuitively, a larger value of α_L makes the interval of service starvation of the light queue following the arrival of a large job into the heavy queue shorter, thus improving the response time tail.

Theorem 3. *Suppose that the arrival rates lie in the subset Λ' of the stability region. Then under max-weight-α scheduling between queues with $\alpha_L > \alpha_H = 1$, and First-Come-First-Served scheduling within the light queue,*

$$\Gamma(V_L) = \lim_{x\to\infty} -\frac{\log P\left(V_L > x\right)}{\log(x)} = \alpha_L(\theta_H - 1).$$

Theorem 4. *Suppose that the arrival rates lie in the subset Λ' of the stability region. Then under max-weight-α scheduling between queues with $\alpha_L > \alpha_H = 1$, and Preemptive-Last-Come-First-Served scheduling within the light queue,*

$$\bar{\Gamma}(V_L) = \limsup_{x\to\infty} -\frac{\log P\left(V_L > x\right)}{\log(x)} \leq \theta_H - \frac{1}{\alpha_L}.$$

Theorem 3 states that with FCFS scheduling within the light queue, the response time tail index increases linearly with α_L. This means that while the response time distribution in the light queue remains heavy-tailed for all α_L, its tail index can be made arbitrarily large by setting α_L to a large enough value, i.e., by giving the light queue sufficient priority. In contrast, Theorem 4 states that under PLCFS scheduling in the light queue, the tail index remains bounded above by θ_H for all values of α_L. This highlights the importance of choosing the correct intra-queue scheduling policy in order to exploit the priority awarded to it by the inter-queue scheduling policy.

The performance of the heavy queue: Next, we turn to the response time tail in the heavy queue under max-weight-α inter-queue scheduling. The following theorems summarize our results for FCFS and PLCFS intra-queue scheduling in the heavy queue.

Theorem 5. *Under max-weight-α scheduling between queues with $\alpha_L \geq \alpha_H = 1$, and First-Come-First-Served scheduling within the heavy queue,*

$$\Gamma(V_H) = \lim_{x\to\infty} -\frac{\log P\left(V_H > x\right)}{\log(x)} = \theta_H - 1.$$

Theorem 6. *In the wireline scenario, under max-weight-α scheduling policy between queues with $\alpha_L > \frac{\theta_H}{\theta_H - 1}$ and $\alpha_H = 1$, and Preemptive-Last-Come-First-Served scheduling within the heavy queue,*

$$\Gamma(V_H) = \lim_{x\to\infty} -\frac{\log P\left(V_H > x\right)}{\log(x)} = \theta_H. \tag{2}$$

Theorem 5 implies that with FCFS scheduling within the heavy queue, the response time tail index is insensitive to α_L; i.e., it is insensitive to the level of relative priority awarded to the light queue. Moreover, the response time tail index is the same as it would be in an isolated $Geo/GI/1$ queue with the same arrival process as the heavy queue.[2]

With PLCFS scheduling within the heavy queue, we are only able to analyze the wireline scenario, when the inter-queue priority to the light queue being sufficiently high (specifically, $\alpha_L > \frac{\theta_H}{\theta_H - 1}$). For this case, Theorem 6 implies as before that the response time tail index for the heavy queue is insensitive to α_L, and is the same as it would be in an isolated $Geo/GI/1$ queue with the same arrival process.[2] Furthermore, this response time tail index is optimal, since the response time tail index is bounded above by the tail index of the job size distribution (i.e., θ_H). We conjecture that Equation (2) holds even in our general 'wireless' scenario, in which the two queues have a stochastic connectivity with the server.[3]

To summarize, under max-weight-α scheduling, the light queue necessarily experiences heavy-tailed response times. However, by setting α_L large enough, i.e., by awarding sufficiently high priority to the light queue, its response time tail index can be made arbitrarily large, with the correct choice of intra-queue scheduling policy. Further, our results suggest that the response time tail index of the heavy queue is unaffected in this process, and behaves like the response time tail index in an isolated $Geo/GI/1$ queue (with the same arrival process).

Ultimately however, from a fairness standpoint, it is desirable that response times in the light queue are light-tailed. Since the level of priority awarded to the light queue by the max-weight-α policy is insufficient for this to happen, we now analyze the log-max-weight inter-queue policy, which awards an even higher degree of priority to the light queue.

B. Log-max-weight scheduling

In this section, we study the tail behavior of the (stationary) response time distribution in the light queue and the heavy queue under the log-max-weight inter-queue scheduling policy.

The performance of the light queue: Our main result in this section is that under log-max-weight scheduling between queues, and FCFS scheduling within the light queue, the light queue experiences light-tailed response times.

Theorem 7. *Suppose that the arrival rates lie in the subset Λ' of the stability region. Then under log-max-weight scheduling between queues, and First-Come-First-Served scheduling within the light queue, V_L is light-tailed.*

The above theorem implies that *the log-max-weight policy indeed provides sufficient priority to the light queue to make its response time distribution light-tailed*. However, for this to happen, the intra-queue scheduling policy cannot be chosen arbitrarily. In fact, as the following theorem shows, with PLCFS scheduling within the light queue, its response time distribution remains heavy-tailed.

Theorem 8. *Suppose that the arrival rates lie in the subset Λ' of the stability region. Then under log-max-weight schedul-*

[2]In a $Geo/GI/1$ queue with the same arrival process as the heavy queue, it is well known that the response time tail index equals $\theta_H - 1$ under FCFS scheduling, and θ_H under PLCFS scheduling (for example, see [19]).

[3]For a discussion on what makes the extension to the wireless case difficult, see [20, Chap. 5].

ing between queues, and Preemptive-Last-Come-First-Served scheduling within the light queue, V_L is heavy-tailed.

This extreme contrast between the two policies highlights once again the importance of correctly choosing the intra-queue scheduling policy to exploit the priority awarded to the light queue by the inter-queue scheduling policy. Theorems 7 and 8 demonstrate a remarkable phenomenon: with the same service process for the light queue, one intra-queue scheduling discipline results in heavy-tailed response times, whereas another leads to light-tailed response times. In the context of the $Geo/G/1$ queue, the impact of the intra-queue policy is nowhere near this extreme, which highlights how crucial the choice is for the multi-queue setting.

The proof of Theorem 7 relies crucially on the following.

Theorem 9. *Under log-max-weight scheduling between queues, q_L is light-tailed.*

This statement was originally conjectured in [17], but proved only for the wireline scenario. In Section IV, we give a novel proof of Theorem 9 based on Lyapunov arguments.

The performance of the heavy queue: Under log-max-weight inter-queue scheduling, we are only able to analyze the response time tail for the heavy queue in the wireline scenario. For this case, we prove that with both FCFS and PLCFS intra-queue scheduling, the response time distribution has the same tail index as in an isolated $Geo/GI/1$ queue with the same arrival process. These results show that in the wireline scenario, the response time tail index is unaffected by the priority given to the light queue by the log-max-weight policy. We conjecture that the same is true in our general 'wireless' model.[3] The following theorems summarize our results.

Theorem 10. *In the wireline scenario, under log-max-weight scheduling between queues, and First-Come-First-Served scheduling within the heavy queue,*

$$\lim_{x\to\infty} -\frac{\log P\left(V_H > x\right)}{\log(x)} = \theta_H - 1.$$

Theorem 11. *In the wireline scenario, under log-max-weight scheduling between queues, and Preemptive-Last-Come-First-Served scheduling within the heavy queue,*

$$\lim_{x\to\infty} -\frac{\log P\left(V_H > x\right)}{\log(x)} = \theta_H.$$

To summarize, our results show that it is possible to achieve light-tailed response times in the light queue using log-max-weight inter-queue scheduling. In other words, it is possible to design inter-queue and intra-queue scheduling policies for our system such that we maintain throughput optimality, and achieve light-tailed delays for the light queue. Importantly, our results suggest that this can be done without affecting the response time tail index for the heavy queue.

IV. Selected proofs

This section is devoted to proofs of the results presented in Section III. Due to space constraints, we are forced to omit several proofs. In particular, we omit the proofs of all results on the response time tail in the heavy queue, and present only representative proofs of the results on the response time tail in the light queue. We refer the reader to [20, Chap. 5] for all excluded proofs.

We first introduce some notation that is used heavily in our proofs. For functions $\varphi(x)$ and $\xi(x)$, the notation $\varphi(x) \sim \xi(x)$ means $\lim_{x\to\infty} \frac{\varphi(x)}{\xi(x)} = 1$. For $t_1, t_2 \in \mathbb{N}$, $A_L(t_1,t_2) := \sum_{t=t_1}^{t_2} L(t)$, and $A_H(t_1,t_2) := \sum_{t=t_1}^{t_2} H(t)$. Note that $A_L(t_1,t_2)$ and $A_H(t_1,t_2)$ denote, respectively, the number of packets entering the light queue and the heavy queue in slots t_1 through t_2. For $y \in \mathbb{N}$, $A_L^{(y)}(t_1,t_2) := \sum_{t=t_1}^{t_2} L(t)\mathbf{1}_{\{L(t)\le y\}}$. Note that $A_L^{(y)}(t_1,t_2)$ is the number of packets entering the light queue from jobs of size $\le y$ in slots t_1 through t_2. Let $\lambda_L^{(y)} := \mathbb{E}\left[L(1)\mathbf{1}_{\{L(1)\le y\}}\right]$. It follows from the monotone convergence theorem that $\lim_{y\to\infty} \lambda_L^{(y)} = \lambda_L$. Finally, define $\bar{S}_L(t_1,t_2) := \sum_{t=t_1}^{t_2} \mathbf{1}_{\{\eta_L(t)=1,\eta_H(t)=0\}}$. $\bar{S}_L(t_1,t_2)$ is the number of exclusive slots available to the light queue in slots t_1 through t_2. Note that in the wireline scenario, $\bar{S}_L(t_1,t_2) = 0$.

This section is organized as follows. First, we present the proof of Theorem 2, which gives an upper bound on the response time tail index in the light queue under max-weight-α inter-queue scheduling. Next, we prove Theorem 7, which states that under log-max-weight inter-queue scheduling and FCFS scheduling within the light queue, V_L is light-tailed. Finally, we give the proof of Theorem 8, which states that under log-max-weight inter-queue scheduling and PLCFS scheduling within the light queue, V_L is heavy-tailed.

A. Proof of Theorem 2

Our proof of Theorem 2 is based on formalizing the intuition that if a job of size $\Theta(x^{\alpha_L})$ arrives into the heavy queue early in the busy period, then with high probability, the light queue is denied service for a period of $\Omega(x)$ slots, except in its exclusive slots.

The proof relies on the following representation for the response time tail. Consider a tagged busy period of the system. Let N_L denote the number of jobs entering the light queue in this busy period, and $V_{L,i}$, for $i = 1, 2, \cdots, N_L$, denote the response time of the i'th arriving job. The tail of V_L has the following well-known representation.

$$P\left(V_L > x\right) = \frac{\mathbb{E}\left[N_L^{(x)}\right]}{\mathbb{E}\left[N_L\right]}, \tag{3}$$

where $N_L^{(x)} := \sum_{i=1}^{N_L} \mathbf{1}_{\{V_{L,i}>x\}}$ is the number of jobs in the light queue that experience a response time exceeding x in the busy period. The proof proceeds by defining a 'bad' event $I(x)$ such that the bound

$$P\left(V_L > x\right) \ge \frac{P\left(I(x)\right)\mathbb{E}\left[N_L^{(x)} \mid I(x)\right]}{\mathbb{E}\left[N_L\right]} \tag{4}$$

leads us to the statement of the theorem.

Without loss of generality, assume that the busy period under consideration starts in time slot 1. Recall that over the subset Λ' of the stability region, $\bar{p}_L < \lambda_L$, and $\lim_{y\to\infty} \lambda_L^{(y)} = \lambda_L$. Pick y large enough so that $\bar{p}_L < \lambda_L^{(y)}$. Let $\delta := (\lambda_L^{(y)} - \bar{p}_L)/4$.

We are now ready to define the event $I(x)$. Fix $\epsilon > 0$.

$$\begin{aligned} I(x) \quad := \quad & \left\{H(1) > \left\lceil \frac{xy}{\delta} \right\rceil + (\lambda_L + \epsilon)^{\alpha_L} \left\lceil \frac{xy}{\delta} \right\rceil^{\alpha_L} \right\} \bigcap \\ & \left\{A_L\left(1, \left\lceil \frac{xy}{\delta} \right\rceil\right) < (\lambda_L + \epsilon) \left\lceil \frac{xy}{\delta} \right\rceil \right\} \bigcap \\ & \left\{\bar{S}_L\left(1, \left\lceil \frac{xy}{\delta} \right\rceil\right) < (\bar{p}_L + \delta) \left\lceil \frac{xy}{\delta} \right\rceil \right\} \bigcap \\ & \left\{A_L^{(y)}\left(1, \left\lceil \frac{xy}{\delta} \right\rceil\right) > (\lambda_L^{(y)} - \delta) \left\lceil \frac{xy}{\delta} \right\rceil \right\} \\ & =: I_1(x) \cap I_2(x) \cap I_3(x) \cap I_4(x). \end{aligned}$$

Informally, the event $I_1(x)$ corresponds to the busy period starting with a 'large' job of size $O(x^{\alpha_L})$ entering the heavy queue. The events $I_2(x)$, $I_3(x)$, and $I_4(x)$ state that the number of packet arrivals into the light queue and number of exclusive slots for the light queue over the interval from slot 1 to slot $\left\lceil \frac{xy}{\delta} \right\rceil$ do not deviate much from their 'law of large numbers' estimates. Indeed, the weak law of large numbers implies that the events $I_2(x)$, $I_3(x)$, and $I_4(x)$ have a probability approaching 1 as $x \to \infty$.

Next, we show that the event $I(x)$ implies that at least x jobs entering the light queue in the busy period under consideration experience a response time exceeding x time slots. To see this, note that the event $I_1(x) \cap I_2(x)$ implies that the heavy queue has priority over the light queue in slots 1 through $\left\lceil \frac{xy}{\delta} \right\rceil$. Indeed, $I_1(x)$ implies that the length of the heavy queue remains greater than $(\lambda_L + \epsilon)^{\alpha_L} \left\lceil \frac{xy}{\delta} \right\rceil^{\alpha_L}$ over this interval, and $I_2(x)$ implies that the length of the light queue never exceeds $(\lambda_L + \epsilon) \left\lceil \frac{xy}{\delta} \right\rceil$ over the same interval. As a result, under event $I(x)$, the light queue receives service only in its exclusive slots until time $\left\lceil \frac{xy}{\delta} \right\rceil$. Note that $I_3(x)$ gives an upper bound on the number of exclusive slots received by the light queue until time $\left\lceil \frac{xy}{\delta} \right\rceil$. Finally, $I_4(x)$ gives a lower bound on the number packets arriving into the light queue until time $\left\lceil \frac{xy}{\delta} \right\rceil$ from jobs of size $\leq y$. Therefore, under event $I(x)$, the number of packets remaining in the light queue after time slot $\left\lceil \frac{xy}{\delta} \right\rceil$, corresponding to jobs of size $\leq y$ exceeds

$$(\lambda_L^{(y)} - \delta) \left\lceil \frac{xy}{\delta} \right\rceil - (\bar{p}_L + \delta) \left\lceil \frac{xy}{\delta} \right\rceil = 2\delta \left\lceil \frac{xy}{\delta} \right\rceil \geq 2xy.$$

Now, since the corresponding jobs have a size of at most y, we conclude that under $I(x)$, the light queue contains at least $2x$ jobs at the end of $\left\lceil \frac{xy}{\delta} \right\rceil$ slots. Since each of these jobs requires at least one slot of service to complete, we conclude that under $I(x)$, at least $x-1$ jobs experience a response time exceeding x in the busy period under consideration.

Returning now to the bound (4), we have defined the event $I(x)$ such that $\mathbb{E}\left[N_L^{(x)} \mid I(x)\right] \geq x - 1$. To bound the probability of $I(x)$, note that

$$P\left(I(x)\right) = P\left(I_1(x)\right) \; P\left(I_2(x) \cap I_3(x) \cap I_4(x)\right),$$

since the arrival process into the heavy queue is independent of the arrival process into the light queue and the queue connectivity processes. Invoking the weak law of large numbers, we conclude that for $\nu \in (0,1)$, $P\left(I_2(x) \cap I_3(x) \cap I_4(x)\right) > (1-\nu)$ for large enough x. Therefore, for large enough x,

$$P\left(V_L > x\right) \geq \frac{1-\nu}{\mathbb{E}\left[N_L\right]}(x-1) \; P\left(I_1(x)\right).$$

The above statement implies that

$$\begin{aligned} \limsup_{x\to\infty} -\frac{\log P\left(V_L > x\right)}{\log(x)} &\leq \lim_{x\to\infty} -\frac{\log P\left(I_1(x)\right)}{\log(x)} - 1 \\ &= \alpha_L \theta_H - 1, \end{aligned}$$

where the last step above uses the fact that $H(1) \in \mathcal{RV}(\theta_H)$, which implies that

$$\lim_{x\to\infty} -\frac{\log P\left(H(1) > \left\lceil \frac{xy}{\delta} \right\rceil + (\lambda_L + \epsilon)^{\alpha_L} \left\lceil \frac{xy}{\delta} \right\rceil^{\alpha_L}\right)}{\log(x)} = \alpha_L \theta_H.$$

This completes the proof.

B. Proof of Theorem 7

When proving Theorem 7 we restrict ourselves to the 'wireless' case, i.e., $p_L, p_H \in (0,1)$, in this paper. The proof for the wireline case is more involved, and can be found in [20, Chap. 5]. Our proof relies crucially on Theorem 9, which we prove first.

The proof of Theorem 9 utilizes a property (Lemma 12 below) of the class of long-tailed distributions, which is an important subclass of heavy-tailed distributions. Formally, a non-negative random variable X (or its d.f. F_X) is said to be *long-tailed* (denoted $X \in \mathcal{L}$) if $\lim_{x\to\infty} \frac{P(X>x+y)}{P(X>x)} = 1$ for all $y > 0$. The class of regularly varying distributions is a strict subset of the class of long-tailed distributions, which in turn is a strict subset of the class of heavy-tailed distributions [21]. We use the following sufficient condition for a distribution to be long-tailed (see [20, Chap. 5] for the proof).

Lemma 12. *Suppose X is a non-negative random variable. If $\Psi_X(x)$ is non-increasing with*

$$\lim_{x\to\infty} \Psi_X(x) = 0,$$

then $X \in \mathcal{L}$.

Additionally, the proof of Theorem 9 relies on the following lemma.

Lemma 13. *Suppose that F_X is the distribution function corresponding to a non-negative random variable. If $F_X \in \mathcal{L}$, and $\bar{F}_X(x) := 1 - F_X(x)$ is strictly decreasing over $x \geq 0$, then*

$$\mathbb{E}\left[\frac{1}{\bar{F}_X(q_L)}\right] < \infty \text{ and } \mathbb{E}\left[\frac{1}{\bar{F}_X(\log(1+q_H))}\right] < \infty.$$

The above lemma is a direct consequence of Theorem 1 in [6]. Note that if $F_X \in \mathcal{L}$, then $1/\bar{F}(x)$ grows sub-exponentially. Therefore, Lemma 13 states that certain *sub-exponential* moments of q_L are finite. However, in order to prove that q_L is light-tailed, we need to show that certain *exponential* moments of q_L are finite, i.e., $\mathbb{E}\left[e^{\beta q_L}\right] < \infty$ for some $\beta > 0$. We do this as follows.

Proof of Theorem 9: For the purpose of obtaining a contradiction, let us assume that q_L is heavy-tailed. Invoking Lemma 1, we conclude that $\liminf_{x\to\infty} \Psi_{q_L}(x) = 0$. Fix $\delta \in (0,1)$. It is easy to see that there exists a strictly increasing integer sequence $\{x_k\}_{k\geq 1}$, with $x_1 = 0$, and $x_k \overset{k\uparrow\infty}{\to} \infty$ such that

(*i*) $\Psi_{q_L}(x_k)$ is non-decreasing in k, with $\lim_{k\to\infty} \Psi_{q_L}(x_k) = 0$,
(*ii*) $\bar{F}_{q_L}(x_{k+1}) \leq (1-\delta)\bar{F}_{q_L}(x_k)$ for $k \geq 1$.

We now define a distribution F_Y that agrees with F_{q_L} along the sequence $\{x_k\}$ such that F_Y satisfies the conditions of Lemma 13, implying that $\mathbb{E}\left[1/\bar{F}_Y(q_L)\right] < \infty$. We then show via a direct computation that $\mathbb{E}\left[1/\bar{F}_Y(q_L)\right] = \infty$. This gives us a contradiction, proving that q_L is light-tailed.

We define the distribution F_Y as follows. $\bar{F}_Y(x_k) = \bar{F}_{q_L}(x_k)$ for all $k \geq 1$. For $x \in (x_k, x_{k+1})$,

$$\log(\bar{F}_Y(x)) = \log(\bar{F}_Y(x_k)) + \frac{x - x_k}{x_{k+1} - x_k}\left(\log(\bar{F}_Y(x_{k+1})) - \log(\bar{F}_Y(x_k))\right). \quad (5)$$

Note that for $x \in (x_k, x_{k+1})$, $\log(\bar{F}_Y(x))$ is defined by linearly interpolating between $\log(\bar{F}_Y(x_k))$ and $\log(\bar{F}_Y(x_{k+1}))$. Equation (5) implies, via simple algebraic manipulations that for $x \in (x_k, x_{k+1})$,

$$\begin{aligned}\Psi_Y(x) &= \frac{\log(\bar{F}_Y(x_k)) - \log(\bar{F}_Y(x_{k+1}))}{x_{k+1} - x_k} \\ &\quad + \frac{1}{x}\frac{x_k x_{k+1}\left(\Psi_Y(x_k) - \Psi_Y(x_{k+1})\right)}{x_{k+1} - x_k} \\ &=: \nu_1 + \frac{\nu_2}{x},\end{aligned}$$

where $\nu_1 > 0$, $\nu_2 \geq 0$. This implies that $\Psi_Y(x)$ is non-decreasing over $x \geq 0$, with $\lim_{x\to\infty} \Psi_Y(x) = 0$. From Lemma 12, we conclude that then $F_Y \in \mathcal{L}$. Moreover, since $\bar{F}_Y(x)$ is strictly decreasing over $x \geq 0$ by definition, Lemma 13 implies that $\mathbb{E}\left[1/\bar{F}_Y(q_L)\right] < \infty$.

We now show through a direct computation that $\mathbb{E}\left[1/\bar{F}_Y(q_L)\right] = \infty$. Pick $k_0 \in \mathbb{N}$.

$$\begin{aligned}\mathbb{E}\left[\frac{1}{\bar{F}_Y(q_L)}\right] &\geq \sum_{x=1}^{x_{k_0+1}} \frac{1}{\bar{F}_Y(x)} P\left(q_L = x\right) \\ &= \sum_{k=1}^{k_0}\sum_{x=x_k+1}^{x_{k+1}} \frac{1}{\bar{F}_Y(x)} P\left(q_L = x\right) \\ &\geq \sum_{k=1}^{k_0}\sum_{x=x_k+1}^{x_{k+1}} \frac{1}{\bar{F}_Y(x_k)} P\left(q_L = x\right) \\ &= \sum_{k=1}^{k_0} \frac{\bar{F}_{q_L}(x_k) - \bar{F}_{q_L}(x_{k+1})}{\bar{F}_Y(x_k)}.\end{aligned}$$

Now, since $\bar{F}_Y$ and $\bar{F}_{q_L}$ agree along the sequence $\{x_k\}$,

$$\mathbb{E}\left[\frac{1}{\bar{F}_Y(q_L)}\right] \geq \sum_{k=1}^{k_0} \frac{\bar{F}_Y(x_k) - \bar{F}_Y(x_{k+1})}{\bar{F}_Y(x_k)} \geq \sum_{k=1}^{k_0} \delta = k_0\delta, \quad (6)$$

where the last step above uses the fact that $\bar{F}_Y(x_{k+1}) \leq (1-\delta)\bar{F}_Y(x_k)$ for $k \geq 1$. Since $\mathbb{E}\left[1/\bar{F}_Y(q_L)\right] \geq k_0\delta$ for any $k_0 \in \mathbb{N}$, it follows that $\mathbb{E}\left[1/\bar{F}_Y(q_L)\right] = \infty$. This gives us a contradiction, which proves that q_L is light-tailed. ∎

We are now ready to give the proof of Theorem 7 in the 'wireless' case.

Proof of Theorem 7 for the case $p_L, p_H \in (0,1)$: Consider a tagged job entering the light queue in slot 0 in steady state. The tagged job has size $L(0) > 0$ and sees a queue length $q_L(0)$ in the light queue. Theorem 9 implies that $q_L(0)$ is light-tailed.

Let us denote the response time of the tagged job by V_L. We need to prove that V_L is light-tailed. Note that

$$q_L(1) = q_L(0) - \mathbf{1}_{\{\text{light queue for service in slot 0}\}} + L(0).$$

Since $q_L(0)$ and $L(0)$ are both light-tailed, it follows that $q_L(1)$ is light-tailed. Now, since the light queue uses FCFS scheduling, V_L is simply equal to the time it takes for the light queue to receive service $q_L(1)$ times. Define $T := \min\{x \in \mathbb{N} \mid \bar{S}_L(1,x) \geq q_L(1)\}$. Note that T is the time it takes after slot 0 for the light queue to see $q_L(1)$ exclusive slots. Clearly, $V_L \leq T$. Fix small $\epsilon > 0$. We have

$$\begin{aligned}P(V_L > x) &\leq P\left(T > x\right) \\ &= P\left(\bar{S}_L(1,x) < q_L(1)\right) \\ &= P\left(\bar{S}_L(1,x) < q_L(1);\ q_L(1) > (\bar{p}_L - \epsilon)x\right) \\ &\quad + P\left(\bar{S}_L(1,x) < q_L(1);\ q_L(1) \leq (\bar{p}_L - \epsilon)x\right) \\ &\leq P\left(q_L(1) > (\bar{p}_L - \epsilon)x\right) \\ &\quad + P\left(\bar{S}_L(1,x) < (\bar{p}_L - \epsilon)x\right) =: I + II.\end{aligned}$$

To prove that V_L is light-tailed, it suffices to show that Terms I and II above are both bounded above by exponentially decaying functions of the form $\nu e^{-\phi x}$, for $\mu, \phi > 0$. That this is true of Term I follows from the fact that $q_L(1)$ is light-tailed. That Term II is similarly bounded follows from a Chernoff bound. This completes the proof. ∎

C. Proof of Theorem 8

As in the proof of Theorem 2, our proof of Theorem 8 is based on defining a 'bad' event $I(x)$ in a tagged busy period, such that the bound (4) leads us to the statement of the theorem. Informally, the event $I(x)$ involves a large enough job arriving into the heavy queue to start the busy period, resulting in $\Omega(\log(x))$ jobs in the light queue experiencing a response time of $\Omega(x)$ slots in the busy period.

Without loss of generality, assume that the busy period under consideration starts in time slot 1. Recall that over the subset Λ' of the stability region, $\bar{p}_L < \lambda_L$, and $\lim_{y\to\infty} \lambda_L^{(y)} = \lambda_L$. Pick y large enough so that $\bar{p}_L < \lambda_L^{(y)}$. Pick $\delta > 0$ such that $\delta \leq (\lambda_L^{(y)} - \bar{p}_L)/4$.

Our 'bad' event $I(x) := G(x) \cap H(x)$, where we define and interpret the events $G(x)$ and $H(x)$ below. We start with the definition of $G(x)$. This event is parameterized by $\beta \in \mathbb{N}$, whose value we fix later.

$$\begin{aligned}G(x) := &\left\{H(1) > \beta\left\lfloor\log(x)\right\rfloor + x^{\beta(\lambda_L+2\delta)} + x + \beta\delta\log(x)\right\} \\ &\bigcap\left\{A_L\left(1, \beta\left\lfloor\log(x)\right\rfloor\right) < (\lambda_L + \delta)\beta\left\lfloor\log(x)\right\rfloor\right\} \\ &\bigcap\left\{\bar{S}_L\left(1, \beta\left\lfloor\log(x)\right\rfloor\right) < (\bar{p}_L + \delta)\beta\left\lfloor\log(x)\right\rfloor\right\} \\ &\bigcap\left\{A_L^{(y)}\left(1, \beta\left\lfloor\log(x)\right\rfloor\right) > (\lambda_L^{(y)} - \delta)\beta\left\lfloor\log(x)\right\rfloor\right\} \\ =: &\ G_1(x) \cap G_2(x) \cap G_3(x) \cap G_4(x).\end{aligned}$$

Roughly, $G(x)$ states that a job of size $\Theta(x^{\max\{\beta(\lambda_L+2\delta),1\}})$ arrives into the heavy queue at the start of the busy period, and the number of arrivals in the light queue, as well as the number of exclusive slots seen by it in slots 1 through $\beta\left\lfloor\log(x)\right\rfloor$ do not deviate much from their 'law of large numbers' estimates. The following lemma states a key implication of $G(x)$.

Lemma 14. *$G(x)$ implies that at the end of $\beta\lfloor\log(x)\rfloor$ slots, the light queue contains at least $2\beta\delta\lfloor\log(x)\rfloor$ packets from jobs of size $\leq y$.*

We omit the proof of Lemma 14 since it uses arguments very similar to those in Theorem 2. Now, invoking the weak law of large numbers, we know that $P(G_2(x)\cap G_3(x)\cap G_4(x))$ approaches 1 as $x\to\infty$. Therefore, fixing $\nu\in(0,1)$,

$$P(G(x))\geq(1-\nu)P(G_1(x)) \text{ for large enough } x. \quad (7)$$

Next, we define the event $H(x)$. Let

$$n(x):=\left\lceil\frac{x}{\lfloor\beta\delta\lfloor\log(x)\rfloor\rfloor}\right\rceil,\quad m(x):=\lfloor\beta\delta\lfloor\log(x)\rfloor\rfloor.$$

The event $H(x)$ concerns arrivals into the light queue, and exclusive slots available to it over $n(x)m(x)$ slots following slot $\beta\lfloor\log(x)\rfloor$. Specifically, the event $H(x)$ states that the number of arrivals in the light queue, as well as the number of exclusive slots available to it, do not deviate much from the corresponding 'law of large numbers' estimates over $n(x)$ periods, each period being $m(x)$ slots long. For notational convenience, define $t[k]:=\beta\lfloor\log(x)\rfloor+(k-1)m(x)$. Formally,

$$H(x):=\bigcap_{k=1,2,\cdots,n(x)}H_k(x),$$

where

$$\begin{aligned}H_k(x):=&\left\{\bar{S}_L(t[k]+1,t[k+1])<(\bar{p}_L+\delta)m(x)\right\}\bigcap\\&\{A_L(t[k]+1,t[k+1])>(\lambda_L-\delta)m(x)\}\\=:&\,H_{k,1}(x)\cap H_{k,2}(x).\end{aligned}$$

The following lemma states the key implication of our 'bad' event $I(x)=G(x)\cap H(x)$.

Lemma 15. *The event $I(x)=G(x)\cap H(x)$ implies that over $n(x)m(x)$ slots following slot $\beta\lfloor\log(x)\rfloor$, the occupancy of the light queue never dips more than $m(x)$ below its level after slot $\beta\lfloor\log(x)\rfloor$.*

Due to space limitations, we omit the proof of this lemma (see [20, Chap. 5] for the proof). Now, we know from Lemma 14 that under event $I(x)$, at the end of slot $\beta\lfloor\log(x)\rfloor$, there are at least $2\beta\delta\lfloor\log(x)\rfloor$ packets in the light queue from jobs of size $\leq y$. Since the light queue uses PLCFS, we conclude from Lemma 15 that at least $\beta\delta\lfloor\log(x)\rfloor$ packets, from jobs of size $\leq y$, stay in queue for more than $n(x)m(x)$ slots. Since $n(x)m(x)\geq x$, this in turn implies that under event $I(x)$, at least $\frac{2\beta\delta\lfloor\log(x)\rfloor}{y}$ jobs in the light queue experience a response time exceeding x, i.e.,

$$\mathbb{E}\left[N_L^{(x)}\mid I(x)\right]\geq\frac{2\beta\delta\lfloor\log(x)\rfloor}{y}. \quad (8)$$

Note that the Chernoff bound implies that there exists $\tau>0$ such that $P(H_{k,1}(x))\geq 1-e^{-\tau m(x)}$ and $P(H_{k,2}(x))\geq 1-e^{-\tau m(x)}$. Therefore, $P(H_k(x))\geq 1-2e^{-\tau m(x)}$, implying that

$$P(H(x))\geq\left(1-2e^{-\tau m(x)}\right)^{n(x)}.$$

Let us fix $\beta>1/\tau\delta$. For this choice of β, it is easy to show that $P(H(x))\overset{x\uparrow\infty}{\to}1$, implying that

$$P(H(x))\geq(1-\nu) \text{ for large enough } x. \quad (9)$$

Finally, returning to our bound (4), it is easy to show that (7), (8), and (9) imply that $\limsup_{x\to\infty}-\frac{\log P(V_L>x)}{\log(x)}<\infty$, which in turn implies that V_L is heavy-tailed. This completes the proof.

REFERENCES

[1] L. Tassiulas and A. Ephremides, "Dynamic server allocation to parallel queues with randomly varying connectivity," *IEEE Transactions on Information Theory*, vol. 39, no. 2, pp. 466–478, 1993.

[2] ——, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Transactions on Automatic Control*, vol. 37, no. 12, pp. 1936–1948, 1992.

[3] N. McKeown, A. Mekkittikul, V. Anantharam, and J. Walrand, "Achieving 100% throughput in an input-queued switch," *IEEE Transactions on Communications*, vol. 47, no. 8, pp. 1260–1267, 1999.

[4] M. Neely, E. Modiano, and C. Rohrs, "Power and server allocation in a multi-beam satellite with time varying channels," in *Proceedings of IEEE INFOCOM*, 2002.

[5] A. Stolyar, "Maxweight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic," *The Annals of Applied Probability*, vol. 14, no. 1, pp. 1–53, 2004.

[6] A. Eryilmaz, R. Srikant, and J. R. Perkins, "Stable scheduling policies for fading wireless channels," *IEEE/ACM Transactions on Networking*, vol. 13, no. 2, pp. 411–424, 2005.

[7] A. Brzezinski and E. Modiano, "Dynamic reconfiguration and routing algorithms for IP-over-WDM networks with stochastic traffic," *Journal of Lightwave Technology*, vol. 23, no. 10, 2005.

[8] M. Neely, "Delay analysis for max weight opportunistic scheduling in wireless systems," *IEEE Transactions on Automatic Control*, vol. 54, no. 9, pp. 2137–2150, 2009.

[9] M. Neely, E. Modiano, and C. Rohrs, "Dynamic power allocation and routing for time varying wireless networks," in *Proceedings of IEEE INFOCOM*, 2003.

[10] L. Le, K. Jagannathan, and E. Modiano, "Delay analysis of maximum weight scheduling in wireless ad hoc networks," *Proc. of IEEE CISS*, pp. 389–394, 2009.

[11] A. Ganti, E. Modiano, and J. Tsitsiklis, "Optimal transmission scheduling in symmetric communication models with intermittent connectivity," *IEEE Transactions on Information Theory*, vol. 53, no. 3, pp. 998–1008, 2007.

[12] S. Borst, M. Mandjes, and M. van Uitert, "Generalized processor sharing with light-tailed and heavy-tailed input," *IEEE/ACM Transactions on Networking*, vol. 11, no. 5, p. 834, 2003.

[13] S. Borst, O. Boxma, and M. Van Uitert, "The asymptotic workload behavior of two coupled queues," *Queueing Systems*, vol. 43, no. 1, pp. 81–102, 2003.

[14] O. Boxma, Q. Deng, and A. Zwart, "Waiting-time asymptotics for the M/G/2 queue with heterogeneous servers," *Queueing Systems*, vol. 40, no. 1, pp. 5–31, 2002.

[15] M. Markakis, E. Modiano, and J. Tsitsiklis, "Scheduling policies for single-hop networks with heavy-tailed traffic," in *Proceedings of the 47th Annual Allerton Conference on Communication, Control, and Computing*, 2009.

[16] K. Jagannathan, M. Markakis, E. Modiano, and J. Tsitsiklis, "Throughput optimal scheduling in the presence of heavy-tailed traffic," in *Proceedings of the 48th Annual Allerton Conference on Communication, Control, and Computing*, 2010.

[17] K. Jagannathan, "Asymptotic performance of queue length based network control policies," Ph.D. dissertation, Massachusetts Institute of Technology, 2010.

[18] L. Huang, S. Moeller, M. Neely, and B. Krishnamachari, "Lifo-backpressure achieves near optimal utility-delay tradeoff," in *Proceedings of WiOpt*, 2011.

[19] O. Boxma and B. Zwart, "Tails in scheduling," *Performance Evaluation Review*, vol. 34, no. 4, pp. 13–20, 2007.

[20] J. Nair, "Scheduling for heavy-tailed and light-tailed workloads in queueing systems," Ph.D. dissertation, California Institute of Technology, 2012.

[21] K. Sigman, "Appendix: A primer on heavy-tailed distributions," *Queueing Systems*, vol. 33, no. 1, pp. 261–275, 1999.

[22] S. Resnick, *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer, 2007.